CV

27 June 2021 21:42

CROSS - VALIDATION :-

We typically split data into D_{train} & D_{test}.

We train model on D_{train} & we get the accuracy on D_{test}.

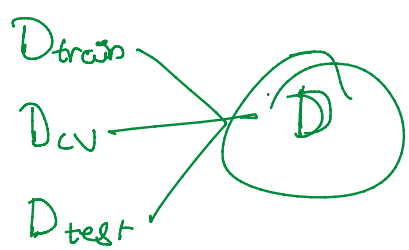
But we can't generalize this result on future data.

Because to evaluate / to do hyper parameter tuning we have used D_{test} here.

So we can't generalize the result on future unseen data.

So we need to have some other idea.

We can split our data into



D_{cv} → cross validation data.

D_{train} → Train the model.

D_{cv} → Hyper parameter tuning

D_{test} → Used as unseen data for generalization.

By splitting train Data into train & CV, we loose [illegible] training. Hence for better performance

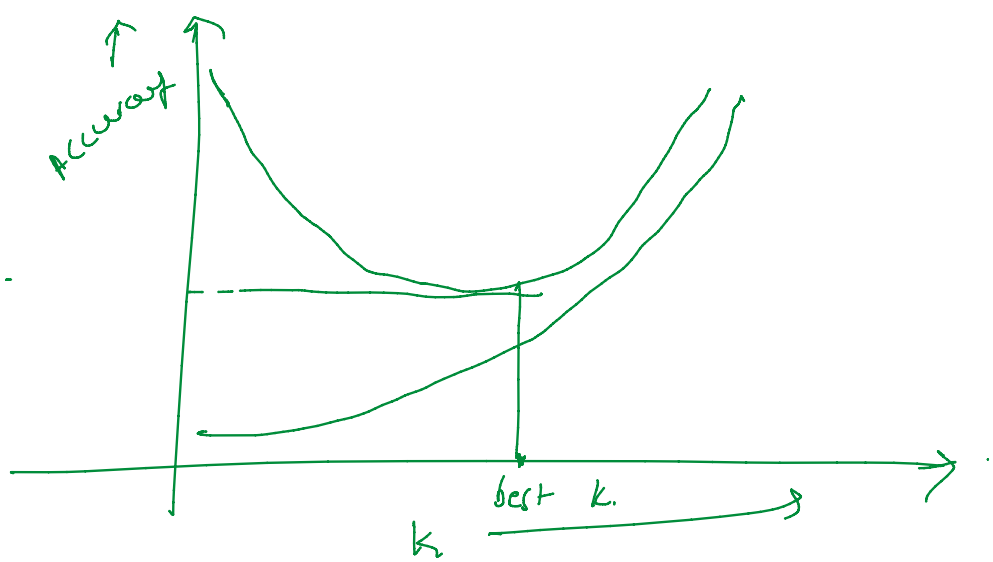
By splitting some data for training. Hence for better performance we use, <u>k'-fold CV</u>

typically k'=10

D_1	D_2	D_3	D_4

	CV	Train	Accuracy
for k=1	D_1	$D_2 D_3 D_4$	A_1
	D_2	$D_1 D_3 D_4$	A_2
	D_3	$D_1 D_2 D_4$	A_3
	D_4	$D_1 D_2 D_3$	A_4
for k=2			
⋮			
for k=n			

Average Accuracy a_1



So, It is best practice to perform cross validation to find best hyperparameter.

best hyperparameter.